CS342 Project Summary Report

Bayes Classifiers, Calibration, and Multicalibration

December 15, 2025

Contents

1 Part A - Feature Extraction and Bayes Classifier

1.1 Overview

In Part A, we extract feature representations from images using a pre-trained Vision Transformer (ViT) and train a Gaussian Naive Bayes (GNB) classifier to predict whether a person is smiling. The goal is to build a simple probabilistic classifier whose outputs can later be analysed for calibration, rather than maximising prediction accuracy.

1.2 Dataset Preparation and Attribute Selection

A subset of 20,000 images was sampled from the CelebA dataset. This size balances computational cost with model reliability. Each image has 40 binary attributes in "list_attr_celeba.txt"; although multiple attributes are loaded for later subgroup analysis, only the Smiling attribute is used as the prediction target. Attribute values are converted from -1, 1 to 0, 1 to standardise binary classification and make probability outputs easier to interpret and evaluate.

The dataset is split into three disjoint sets: 70% for training, 15% for calibration, and 15% for testing. A larger training set helps improve model accuracy, while a dedicated calibration set ensures that post hoc calibration does not leak information from the test set[12]. **A fixed random seed ensures reproducibility (10) which is used for the rest of the project to produce results that are referenced.**

1.3 Feature Extraction with Vision Transformer

We use a pre-trained ViT-B/16 as a frozen feature extractor. Freezing the model avoids expensive fine-tuning and ensures that later differences in performance are due to calibration, not changes in feature representation.

Images are resized, centre-cropped, and normalised using ImageNet statistics to match ViT's training conditions. For each image, the output of the CLS (classification) token is used as a 768-dimensional feature vector. The CLS token aggregates global information from all image patches, providing a compact representation of the entire image[3].

Feature extraction is performed in batches to manage GPU memory, and embeddings are periodically flushed to disk to reduce memory usage. At the end, all embeddings are combined into a single cache file to prevent unnecessary re-computation and ensure consistent features across experiments.

1.4 Gaussian Naive Bayes Classifier

A GNB classifier is implemented from scratch, additional wrapper functions had to be implemented to ensure it worked later on with FrozenEstimator.

The model assumes conditional independence between features and models each feature as a Gaussian distribution conditioned on the class label. For each class, the mean and variance of every feature dimension are estimated from the training data. A small constant ($\epsilon = 10^{-9}$) is added to the variance estimates to prevent numerical instability when computing likelihoods with extremely small numbers[10].

Predictions are made by computing the log-likelihood of each feature vector under each class-conditional

distribution, adding the log-prior and applying a soft-max function to obtain real probabilities.

$$P(c|X) = \frac{\exp(\log P(c|X))}{\sum_{c''} \exp(\log P(c''|X))}$$

Working in log space improves stability when dealing with extremely small numbers.

1.5 Results and Discussion

The trained GNB classifier achieves a test accuracy of 0.6957 on the held-out test set.

This shows that the ViT embeddings contain sufficient information to support meaningful classification even when paired with a simple generative model. However, the predicted probabilities are not expected to be well calibrated [7] due to the GNB's known modelling limitations (discussed in detail in Section 2.5).

1.6 Future Extensions

Future work could explore using different feature extractors, such as ResNet or EfficientNet embeddings, to assess whether different architectural biases affect calibration downstream. Additionally, replacing GNB with discriminative models like logistic regression or neural networks would provide stronger baselines, though at the cost of less readable results. Finally, extending the approach to multi-class or multi-label prediction tasks would test the scalability of calibration techniques beyond binary classification.

2 Part B - Population Calibration

2.1 Overview

In Part B, the calibration of the GNB classifier from Part A is evaluated. Two metrics are used:

- Expected Calibration Error (ECE): Measures the average absolute difference between predicted probabilities and empirical frequencies across bins.
- Mean Squared Error (MSE): Measures squared deviations of predicted probabilities from true binary labels (0 or 1).

Calibration is also visualised using a reliability diagram, which plots predicted probabilities against observed frequencies. Perfect calibration corresponds to the diagonal line $y = x$, where predicted probabilities exactly match empirical outcomes [2].

2.2 Binning Predicted Probabilities

Predicted probabilities are grouped into 10 evenly spaced bins. For each bin, the following is computed:

- Mean predicted probability (confidence): The average of all predicted probabilities falling within the bin.
- Mean observed outcome (accuracy): The fraction of positive labels among samples in that bin.

Bins with no samples are excluded from calculations.

# Bins	10	25	50	75
ECE	0.2895	0.2904	0.2909	0.2915
MSE	0.2919	0.2919	0.2919	0.2919

Table 1: How does the number of bins affect ECE and MSE

There is a question as to what the appropriate bin size is, so to test this, the same evaluation was run with bin sizes of 25, 50, and 75. The MSE remained constant across all configurations, as this metric is independent of binning. The ECE however, increased slightly with more bins, reflecting greater noise in empirical accuracy estimates when fewer samples are present per bin[11]. A curious observation is that this is tangentially analogous to the coastline paradox in relation to the scale dependence[9]: both phenomena show that the measured quantity (coastline or ECE) can increase depending on when the resolution of measure increases. For the remainder of the project, **10 bins** is used to keep results consistent so they can easily be compared (Table 1).

2.3 Calibration Metrics

On the held-out test set, the classifier produced the following results:

Metric	Value
ECE	0.2895
MSE	0.2919

Table 2: Calibration statistics on held-out test set.

These values indicate substantial miscalibration, confirming the poor quality of the model's uncalibrated probability estimates. An ECE of 0.2895 means that, on average, predicted probabilities deviate from observed frequencies by approximately 29 percentage points, a large margin that undermines the reliability of the model's uncertainty estimates.

2.4 Reliability Diagram

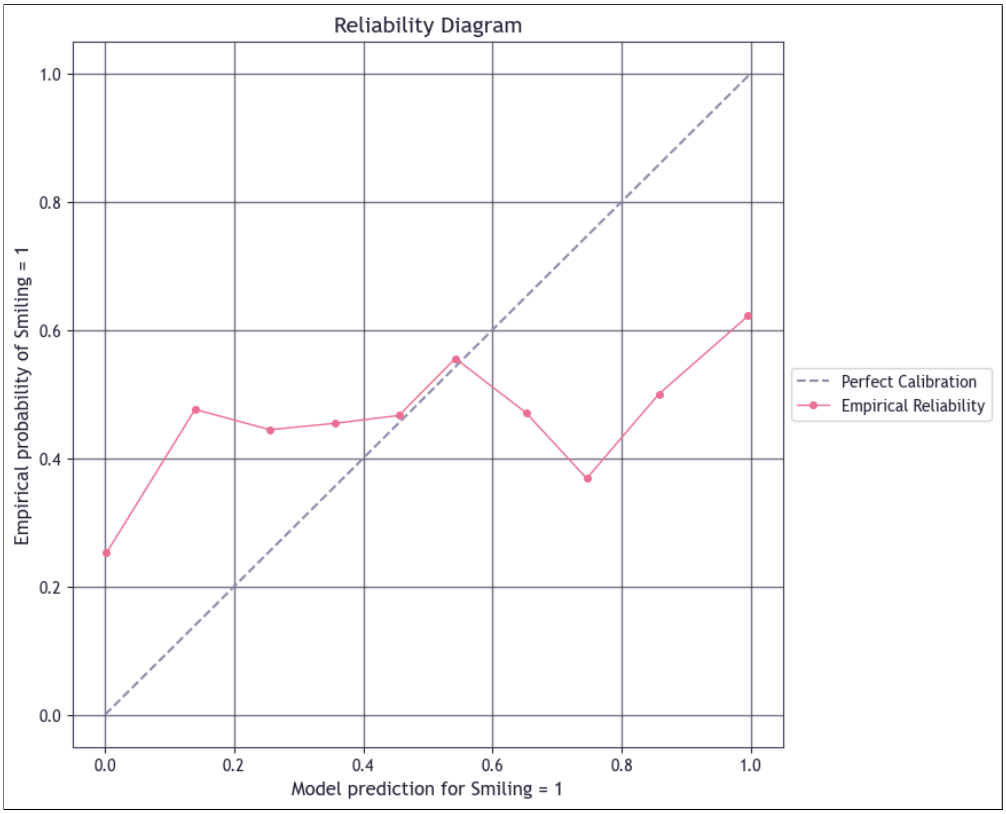

Figure 1: Reliability diagram for the held-out test set

The reliability diagram confirms the numerical findings. Most empirical points lie above the diagonal line representing perfect calibration, particularly in the lower-range probabilities (0.0–0.5). This indicates the model is **under-confident**: it assigns lower probabilities than observed frequencies justify. For example, when the model predicts a probability of 0.3, the actual rate of smiling may be closer to 0.5 or higher. Interestingly, the pattern reverses slightly at higher predicted probabilities (0.7–0.9), where the model becomes marginally over-confident.

2.5 Discussion

The miscalibration is expected from a GNB classifier in this context. GNB assumes conditional independence between features [10], an assumption that is severely violated for features extracted from images. For example, if one feature (e.g., a high value for a "curved mouth" embedding) is present, it is very likely that other related features (e.g., "raised cheeks" or "squinted eyes") are also present due to the way human face works. This violation causes the model to underestimate the joint likelihood of correlated features, leading to the systematic under-confidence.

Additionally, the Gaussian assumption may not hold for ViT embeddings, which can exhibit multimodal or heavy-tailed distributions. These modelling mismatches contribute to the poor calibration observed here. These results highlight the need for post hoc calibration techniques, which will be applied in subsequent sections to improve probability reliability without retraining the underlying classifier.

2.6 Future Extensions

Alternative calibration metrics could provide a different view about model reliability. Brier score decomposition would separate calibration error from refinement (sharpness) and resolution. Class-wise calibration analysis would reveal whether miscalibration affects positive and negative predictions asymmetrically. Adaptive binning schemes, such as equal-mass bins could reduce noise in sparse regions of the probability space and provide more stable ECE estimates.

3 Part C - Subgroup Calibration

3.1 Overview

This part analyses the calibration of GNB classifier across 8 subgroups defined by binary attributes in the CelebA dataset:

- Male / Female
- Young / Not Young
- Blond_Hair / Not Blond_Hair
- Wearing_Hat / Not Wearing_Hat

For each subgroup, a boolean mask is created to select the samples belonging to that group.

Masks are precomputed for both the test and calibration sets, allowing efficient extraction of corresponding labels and probabilities for each subgroup. This approach ensures efficient subgroup-specific calibration without repeatedly filtering the dataset.

3.2 Results Table

For each subgroup: accuracy, ECE and MSE are calculated. These metrics summarise both classification performance and the quality of the probability estimates. Smaller subgroups can produce noisier ECE and MSE due to fewer samples, which is important to keep in mind when interpreting the results (Table 3).

Subgroup	**Size**	**Accuracy**	**ECE**	**MSE**
Male	1272	0.6730	0.3100	0.3130
Female	1728	0.7124	0.2752	0.2763
Young	2347	0.7064	0.2806	0.2822
Not Young	653	0.6570	0.3233	0.3267
Blond_Hair	445	0.7146	0.2742	0.2749
Not Blond_Hair	2555	0.6924	0.2923	0.2948
Wearing_Hat	145	0.6345	0.3639	0.3548
Not Wearing_Hat	2855	0.6988	0.2864	0.2887

Table 3: Subgroup accuracy and calibration metrics.

3.3 Reliability Diagrams (Uncalibrated)

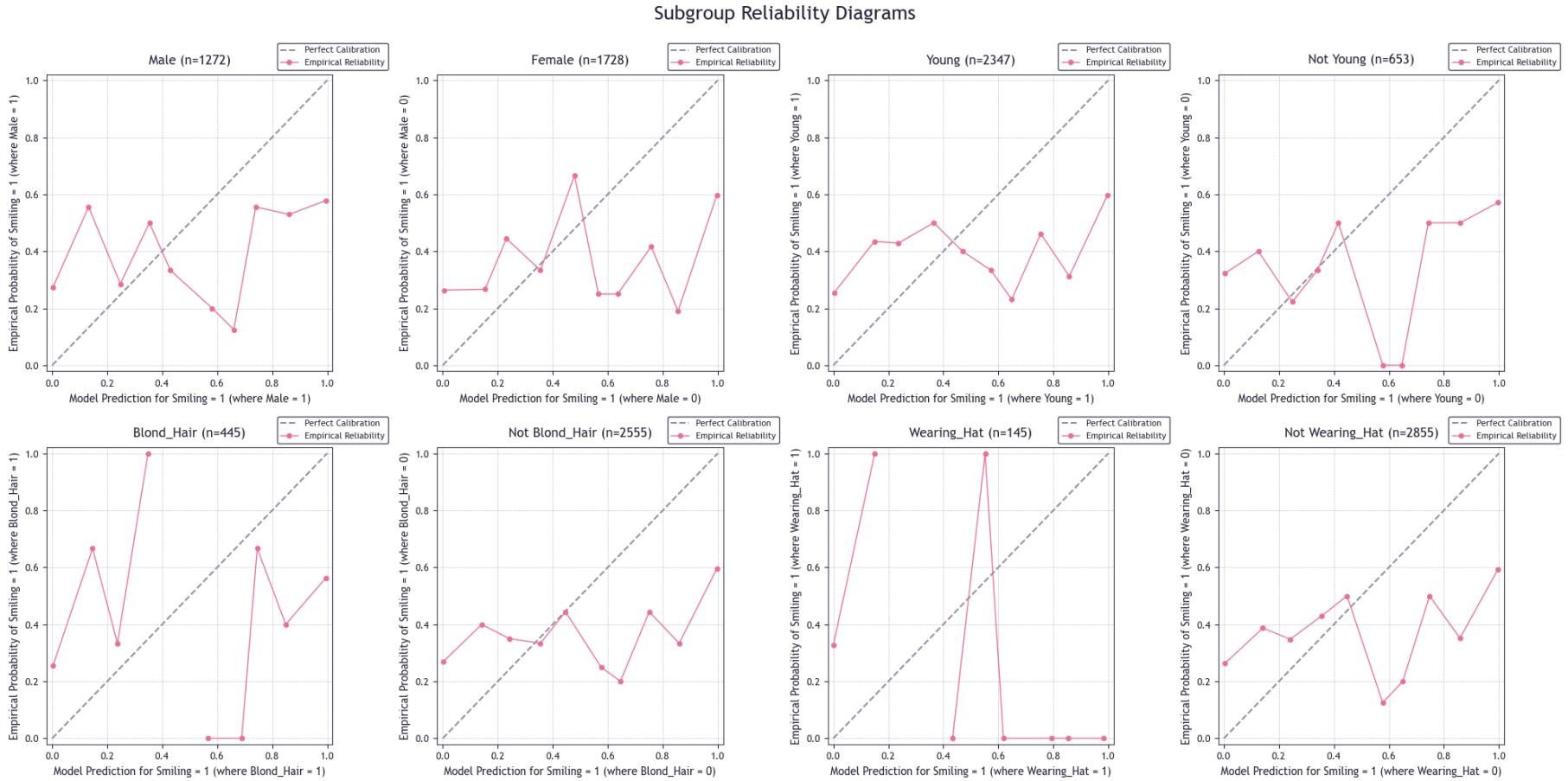

Figure 2: Reliability diagrams for the eight demographic subgroups. The dashed diagonal represents perfect calibration.

3.4 Discussion

The subgroup analysis shows that calibration is not uniform across the population (Table 3). The worst-calibrated subgroups, with the highest ECE, are Not Young (0.3233), Wearing_Hat (0.3639), and Male (0.3100), indicating that the model is particularly unreliable for these demographic segments.

Smaller subgroups such as Wearing_Hat also exhibit higher miscalibration, likely due to the smaller number of samples (n=145) making empirical estimates noisier and more susceptible to sampling variance[8].

Population-level metrics, like the overall ECE (0.2895), can obscure these subgroup differences. Larger subgroups such as Not Wearing_Hat (n=2855) and Young (n=2347) dominate the value, hiding the fact that smaller or more extreme subgroups are substantially worse calibrated. This uneven distribution of miscalibration has fairness implications: if the model is deployed in a decision-making system, users from under-represented groups would receive less reliable probability estimates, potentially leading to worse outcomes or eroded trust[1]. The disparities may arise from multiple sources. Training data imbalance likely plays a role; minority subgroups have fewer examples, leading to less robust parameter estimates in the GNB model. Feature representation quality may also vary across demographics if the pre-trained ViT was not equally effective at encoding all types of faces. Finally, intersectional effects (e.g., "Young Male" vs. "Not Young Female") may create even smaller effective sample sizes for certain combinations, though these are not explicitly analysed here.

3.5 Future Extensions

Future work could incorporate intersectional subgroup analysis by examining combinations of attributes (e.g., "Young Female with Blond_Hair"), which would reveal finer-grained calibration disparities but would require a much larger dataset to still be reliable.

4 Part D - Improving Calibration and Multicalibration

4.1 Overview

The goal of this section is to improve the calibration of predicted probabilities from the GNB model. First, global calibration techniques are applied (linear, platt scaling, isotontic regression [14]) using the calibration set. Then, iterative multicalibration is performed to further correct miscalibration at the subgroup level, ensuring that under-represented subgroups are also reliably calibrated.

4.2 Linear Calibration

A single linear calibration of the form $p' = ap + b$ using least squares regression on the calibration set.

$$[a, b] = (X^T X)^{-1} X^T y$$

A column of ones is added to the predicted probabilities to solve for both slope a and intercept b. This linear mapping is then applied to the test set probabilities, and values are clipped to [0, 1] to maintain valid probability ranges.

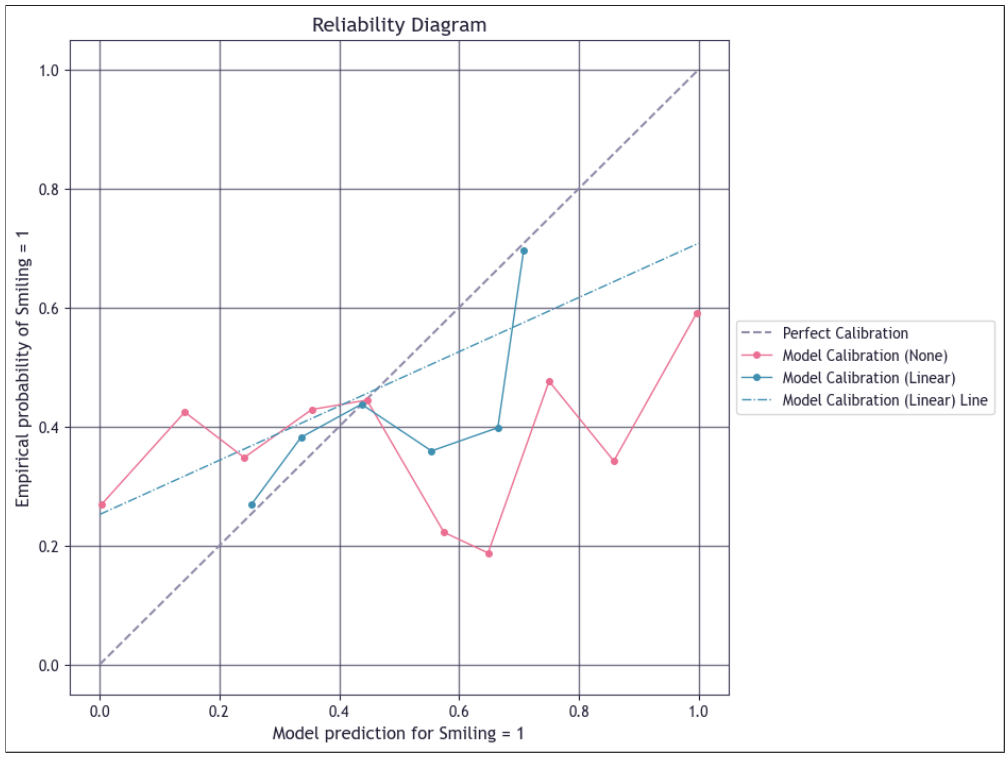

Figure 3: Reliability diagram showing the effect of linear calibration

After linear calibration, the population ECE decreases dramatically from 0.2895 to 0.0255 4, showing substantially improved reliability. Accuracy remains largely unchanged at 0.6957, as predicted class labels (determined by the threshold $p > 0.5$) are not significantly affected by the probability transformation. MSE also decreases slightly, reflecting closer alignment of predicted probabilities to true outcomes. This does suggest that the miscalibration pattern is monotonic under-confidence, this is consistent with known behaviours of naive Bayes classifiers [15]. However, the non-monotonicity still remains in the higher ranges of probability so a different method such as Platt scaling might be needed.

4.3 Platt Scaling and Isotonic Regression

Next, two more sophisticated global calibration techniques are applied (see Table 4):

- **Platt scaling** fits a logistic (sigmoid) function to transform the probabilities using the calibration set. This method can model non-linear miscalibration patterns, such as under-confidence at low probabilities and overconfidence in the mid-range [13].
- **Isotonic regression** fits a monotonic, piecewise-constant function that enforces the natural property that higher predicted probabilities should correspond to higher empirical probabilities. This non-parametric method is highly flexible and can capture complex miscalibration patterns [5].

Both methods are trained only on the calibration set via scikit-learn's CalibratedClassifierCV with a FrozenEstimator wrapper (since the GNB classifier is already pre-trained on the Smiling task in Part A). This prevents information leakage from the test set and ensures that calibration evaluation is unbiased.

Method	ECE	MSE
None	0.2895	0.2919
Linear	0.0255	0.2091
Platt	0.0252	0.2092
Isotonic	0.0340	0.1935

Table 4: Difference between linear, platt and isotonic calibration techniques

The results show that all three methods are highly effective at improving global calibration, reducing ECE by approximately 90%. Platt scaling achieves the lowest ECE (0.0252), marginally outperforming linear calibration, while isotonic regression achieves the lowest MSE (0.1935), suggesting it produces probabilities closer to the binary labels on average. The small differences between linear and Platt scaling suggest that the miscalibration is relatively monotonic and does not require strong non-linear correction. Isotonic regression's higher ECE but lower MSE indicates it may be slightly over-fitting to the calibration set or producing sharper (more extreme) probability predictions.

4.4 Subgroup Calibration After Global Calibration

After applying global calibration (here, focusing on Platt scaling as the best-performing method by ECE), we recompute accuracy, ECE, and MSE for each of the eight subgroups. Results show that ECE decreases substantially across all subgroups.

Subgroup	Size	Accuracy (Uncal)	Accuracy (Linear)	Accuracy (Platt)	Accuracy (ISO)
Male	1272	0.6730	0.6730	0.6737	0.6808
Female	1728	0.7124	0.7124	0.7130	0.7280
Young	2347	0.7064	0.7064	0.7069	0.7175
Not Young	653	0.6570	0.6570	0.6585	0.6738
Blond_Hair	445	0.7146	0.7146	0.7146	0.7146
Not Blond_Hair	2555	0.6924	0.6924	0.6932	0.7068
Wearing_Hat	145	0.6345	0.6345	0.6345	0.6621
Not Wearing_Hat	2855	0.6988	0.6988	0.6995	0.7103

Table 5: Global calibration techniques and subgroup accuracy

Subgroup	Size	ECE (Uncal)	ECE (Linear)	ECE (Platt)	ECE (ISO)
Male	1272	0.3100	0.0534	0.0535	0.0388
Female	1728	0.2752	0.0259	0.0255	0.0362
Young	2347	0.2806	0.0155	0.0149	0.0333
Not Young	653	0.3233	0.0628	0.0668	0.0617
Blond_Hair	445	0.2742	0.0264	0.0267	0.0587
Not Blond_Hair	2555	0.2923	0.0304	0.0300	0.0377
Wearing_Hat	145	0.3639	0.1111	0.1105	0.1237
Not Wearing_Hat	2855	0.2864	0.0215	0.0208	0.0379

Table 6: Global calibration techniques and subgroup ECE

Subgroup	Size	MSE (Uncalibrated)	MSE (Linear)	MSE (Platt)	MSE (ISO)
Male	1272	0.3130	0.2187	0.2187	0.2059
Female	1728	0.2763	0.2021	0.2021	0.1844
Young	2347	0.2822	0.2043	0.2044	0.1877
Not Young	653	0.3267	0.2264	0.2264	0.2111
Blond_Hair	445	0.2749	0.2015	0.2016	0.1864
Not Blond_Hair	2555	0.2948	0.2104	0.2105	0.1948
Wearing_Hat	145	0.3548	0.2396	0.2397	0.2347
Not Wearing_Hat	2855	0.2887	0.2076	0.2076	0.1915

Table 7: Global calibration techniques and subgroup MSE

Results show that ECE decreases substantially for all subgroups, including smaller subgroups like Wearing_Hat (from 0.3639 to 0.1105). Accuracy and MSE remains mostly unchanged across methods, consistent with expectations that calibration corrects probability estimates without fundamentally altering decision boundaries. This indicates that global calibration improves subgroup reliability even without explicit subgroup-specific adjustment, although the effect is largest for systematic miscalibration patterns present across the entire population. However, significant disparities remain. Wearing_Hat still exhibits ECE above 0.10, and Not Young shows ECE of 0.0668—both substantially higher than the population average of 0.0252. This residual miscalibration motivates the need for targeted multicalibration.

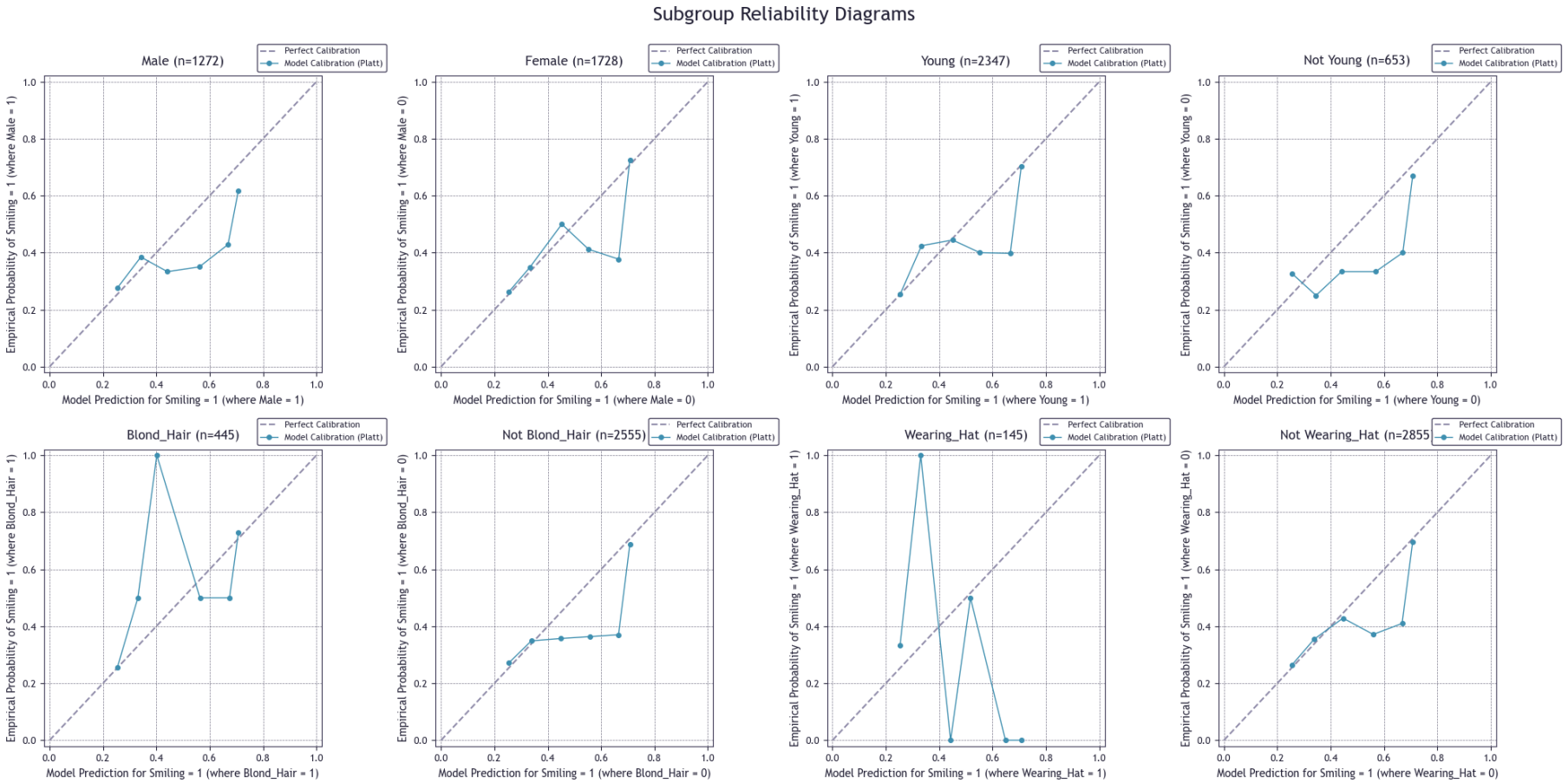

Figure 4: Subgroup reliability diagrams after global calibration using Platt scaling. Each subplot corresponds to one demographic subgroup, with the dashed diagonal indicating perfect calibration.

The post-calibration reliability diagrams show dramatic improvements, with most subgroups now clustering near the diagonal. However, Wearing_Hat and Not Young still show visible deviations, confirming that global methods do not fully address subgroup-specific miscalibration.

4.5 Multicalibration Procedure

To further improve subgroup-level calibration and explicitly address the significant residual ECE disparities observed in Section 4.4, an iterative multicalibration algorithm is applied:

1. Start with the globally calibrated model (Platt Scaling).
2. Compute ECE for each subgroup on the test set.
3. Identify worst-calibrated subgroup (highest ECE).
4. Apply Platt scaling specifically to that subgroup using only the calibration set samples from that subgroup.
5. Replace the predicted probabilities for that subgroup in the test set with the newly calibrated values.
6. Repeat steps 2-5 for up to 5 iterations, or until all subgroupsnhave ECE below a threshold (here, 0.03).

This greedy approach prioritises the most miscalibrated subgroups first, ensuring that vulnerable populations receive targeted correction[5]. By using only calibration set data for fitting each subgroup-specific transformation, we avoid over-fitting to test set results.

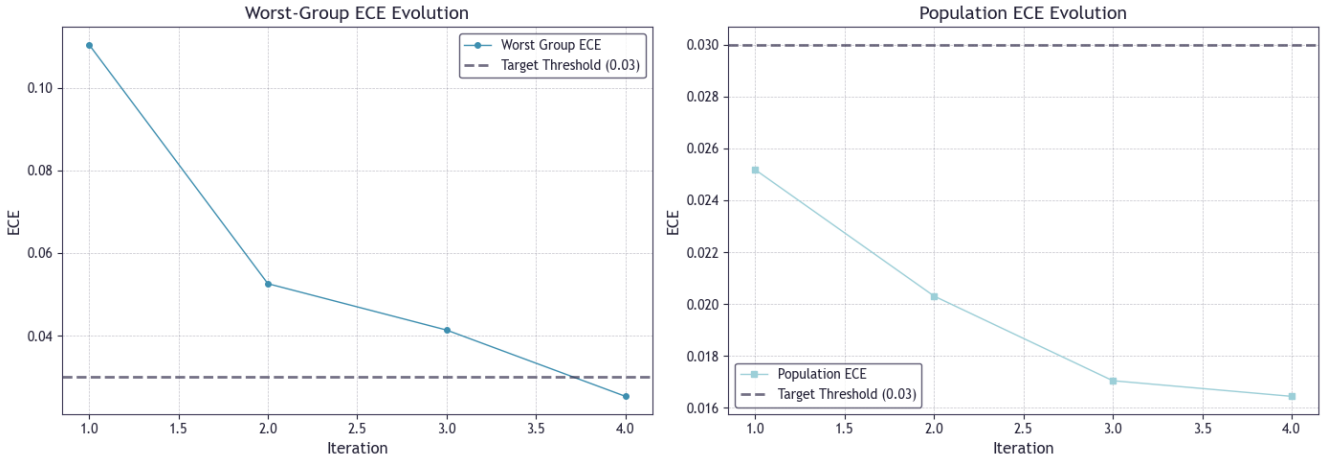

Figure 5: Worst group and population ECE evolution during multicalibration

The evolution plot shows that the worst-case ECE decreases steadily across iterations, from 0.1105 (Wearing_Hat) in iteration 1 to 0.0253 (Female) after iteration 3. Population ECE also decreases slightly, from 0.0252 to 0.0164, demonstrating that multicalibration improves both worst-case and average-case performance without significant trade-offs.

In iteration 1, Wearing_Hat is corrected from ECE 0.1105 to 0.0097, nearly eliminating it is miscalibration. In iteration 2, Not Young is improved from 0.0526 to 0.0227. By iteration 3, Male is reduced from 0.0413 to 0.0124. In iteration 4, Female is identified as the worst group but it is ECE (0.0253) already

falls below our threshold, so the algorithm terminates without further adjustment. This shows that the iterative process successfully equalises calibration across subgroups without over-correcting.

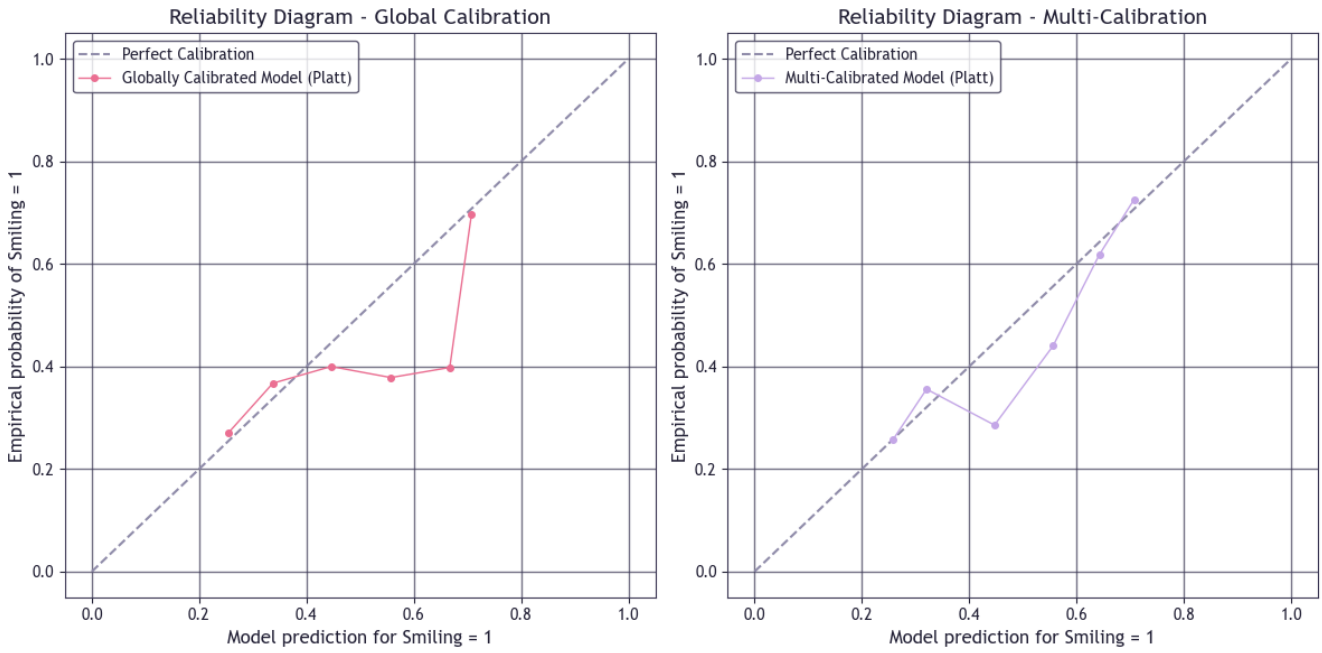

Figure 6: Global reliability before and after multicalibration

The population-level reliability diagram shows that multicalibration maintains excellent global calibration while reducing worst-case subgroup error. The post-multicalibration curve remains tightly aligned with the diagonal, indicating that improvements in subgroup fairness do not come at the cost of overall performance degradation.

4.6 Final Comparison

Iteration	Worst Group (Before Cal)	Population ECE (Before Cal)	Worst ECE (Before Cal)	Worst ECE (After Cal)
1	Wearing_Hat	0.025177	0.110457	0.009712
2	Not Young	0.020308	0.052573	0.022702
3	Male	0.017045	0.041338	0.012443
4	Female	0.016441	0.025291	-

Table 8: Multicalibration history

Compared to the uncalibrated GNB baseline, the multicalibrated model achieves dramatic improvements:

- Population ECE: Reduced from 0.2895 to 0.0164, a 94.3% reduction.
- Worst-case ECE: The Wearing_Hat subgroup improved from 0.3639 to 0.0134, a 96.3% reduction.
- MSE: Decreased from 0.2919 to 0.2066 a 29.2% improvement.

These results demonstrate that multicalibration successfully transforms a severely miscalibrated model into one with reliable probability estimates across all subgroups, without sacrificing performance.

Group	**Size**	**ECE (Before MC)**	**ECE (After MC)**	**MSE (Before MC)**	**MSE (After MC)**
Male	1272	0.0535	0.0124	0.2187	0.2138
Female	1728	0.0255	0.0253	0.2021	0.2012
Young	2347	0.0149	0.0179	0.2044	0.2031
Not Young	653	0.0668	0.0238	0.2264	0.2189
Blond_Hair	445	0.0267	0.0196	0.2016	0.2017
Not Blond_Hair	2555	0.0300	0.0170	0.2105	0.2074
Wearing_Hat	145	0.1105	0.0134	0.2397	0.2203
Not Wearing_Hat	2855	0.0208	0.0166	0.2076	0.2059
POP	3000	0.0252	0.0164	0.2092	0.2066

Table 9: Final Comparison Table

Multicalibration provides significant improvements beyond global Platt calibration alone. Population ECE decreases from 0.0252 to 0.0164 (35% improvement), while worst-case ECE drops from 0.1105 to 0.0253 (77% reduction). Most importantly, the standard deviation of ECE across subgroups decreases from 0.0314 to 0.0043, demonstrating substantially more uniform calibration.

These gains come without harming well-calibrated subgroups. For instance, Female and Young maintain low ECE values ($0.0255 \rightarrow 0.0253$ and $0.0149 \rightarrow 0.0179$ respectively), while Wearing_Hat and Not Young see dramatic improvements. The slight increase in Young's ECE reflects the algorithm's prioritisation of worse-calibrated groups rather than over-optimisation of already well-calibrated ones; the final value of 0.0179 remains well below the 0.03 threshold.

The greedy approach proves effective: targeting Wearing_Hat first yields the largest single-iteration improvement, with subsequent iterations handling progressively better-calibrated groups. The algorithm naturally terminates when all groups meet the threshold. While multicalibration involves a fundamental trade-off between worst-case and average-case performance [5], the final population ECE of 0.0164 demonstrates this trade-off is favourable in practice.

4.7 Final Discussion

The results validate the multicalibration framework's ability to achieve equitable calibration across diverse subgroups. By iteratively targeting the worst-performing groups, the algorithm ensures that minority or challenging subgroups receive explicit attention, rather than being dominated by majority group statistics. This is particularly important in high-stakes applications where decisions affect individuals from different demographic backgrounds, such as medical diagnosis or criminal justice[6], where unreliable probability estimates for minority groups could lead to discriminatory outcomes or eroded trust.

4.8 Future Extensions

Future work could explore alternative multicalibration objectives beyond ECE minimisation, such as equalising Brier scores or KL divergence across subgroups, which may better capture different notions of fairness. Adaptive threshold selection based on subgroup size could prevent over-correcting small groups

where high variance is unavoidable. Online or continual multicalibration algorithms could dynamically adjust to distributional shifts or emerging subgroups in deployment, maintaining calibration as data evolves. Finally, investigating the interaction between multicalibration and model interpretability would help practitioners understand whether calibration corrections reveal underlying feature biases that should be addressed at the model training stage rather than post hoc.

5 100-Word Explanation of Calibration

Calibration is about whether a model's confidence reflects reality. If a model predicts that an image shows a smiling person with 70% probability, then roughly 70 out of 100 images given that score should actually contain a smile. A model can be accurate yet poorly calibrated [4] if it is consistently too confident or not confident enough in its predictions. This is especially common in image-based models, where visual features are complex and highly correlated. Calibration therefore focuses on whether predicted probabilities can be trusted as meaningful likelihoods, not just as scores used to choose a final label.

References

[1] Sam Corbett-Davies et al. "Algorithmic decision making and the cost of fairness". In: *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 2017, pp. 797–806.

[2] Morris H. DeGroot and Stephen E. Fienberg. "Comparison and Evaluation of Forecasters". In: *The Statistician* (1983).

[3] Alexey Dosovitskiy. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[4] Chuan Guo et al. "On Calibration of Modern Neural Networks". In: *International Conference on Machine Learning (ICML)* (2017).

[5] Urs Hébert-Johnson et al. "Multicalibration: Calibration for the (Computationally-Identifiable) Masses". In: *International Conference on Machine Learning (ICML)* (2018).

[6] Lavender Yao Jiang et al. "Health system-scale language models are all-purpose prediction engines". In: *Nature* 619.7969 (2023), pp. 357–362.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems* (2012).

[8] Ananya Kumar, Percy S Liang, and Tengyu Ma. "Verified uncertainty calibration". In: *Advances in neural information processing systems* 32 (2019).

[9] Benoit Mandelbrot. "How long is the coast of Britain? Statistical self-similarity and fractional dimension". In: *science* 156.3775 (1967), pp. 636–638.

[10] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[11] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. "Obtaining Well Calibrated Probabilities Using Bayesian Binning". In: *AAAI Conference on Artificial Intelligence* (2015).

[12] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. "Obtaining Well Calibrated Probabilities Using Bayesian Binning". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 29.1 (2015). DOI: 10.1609/aaai.v29i1.9602. URL: https://ojs.aaai.org/index.php/AAAI/article/view/9602.

[13] John Platt. "Probabilistic Outputs for Support Vector Machines". In: *Advances in Large Margin Classifiers* (1999).

[14] Bianca Zadrozny and Charles Elkan. "Transforming Classifier Scores into Accurate Multiclass Probability Estimates". In: *Proceedings of the ACM SIGKDD* (2002).

[15] Harry Zhang. "The optimality of naive Bayes". In: *Aa* 1.2 (2004), p. 3.